Towards an Abstractive Opinion Summarisation of Multiple Reviews in the Tourism Domain

Cyril Labbé and François Portet

Laboratoire d'Informatique de Grenoble, UJF/Grenoble-INP/CNRS 5217, 38041 Grenoble, France
first.last@imag.fr

Abstract. Since the arrival of Web 2.0, there is an increasing amount of on-line Reviews and Ratings about diverse products or services. The reviews contain general comments as well as highly personal elements or opinions about the customers' experience with the product. Other customers or companies are facing the problem of extracting the relevant information from this mass of reviews. In this paper, we present a comparative study of three different summarisation techniques for reviews analysis. From this study, we propose a general architecture which relies on a customisable abstractive summarisation approach making use of domain knowledge and temporal analysis. The paper ends by identifying research directions for improving the efficiency of review summarisation methods.

Keywords: Review summarisation, Opinion mining, Natural language generation.

1 Introduction

Since the arrival of Web 2.0, costumers of any kind of products or services produce a large amount of on-line Reviews, almost only present as text, and Ratings as ordinal variable. While these reviews have largely contributed to the success of the e-commerce, the problem for a costumer to construct her/his own opinion and to make an informed decision is to make sense of this mass of reviews that contains not only general comments about product features but also highly idiosyncratic information such as opinions or sentiments. Reviews are not only useful for potential costumers but also represent precious information for companies about their own products. A major challenge for society is to make possible an automatic analysis of sets of reviews in order to produce a coherent summary that can be quickly and easily assimilated by humans.

In this paper, we study the problem of review summarisation in the accommodation domain. Automatic summarisation is the process of drawing out the most relevant information from a source to produce a condensed version sometimes biased towards particular users and tasks. Summarisation approaches are generally categorised as: *extractive* when content reduction is addressed by selection or *abstractive* when compression is done by generalisation of what is relevant in the source [1]. While summarisation of technical structured contents is a well implanted technique in industry, summarisation of reviews is a much more recent trend. In this context, the task must face poorly structured contents from a large number of authors (e.g., age, sex, literacy level, etc.) full of

subjective matters expressed via opinion, metaphor, or cultural references. The excerpts (table 1) from an existing database illustrate the variety of reviews for a same hotel.

```
The hotel was well serviced by friendly staff. Great bathroom with a flat floor
shower..... no bath! Mini kitchenette was really handy, great not to have to use
the vanity basin as a kitchen sink! While there are no retail shops close by,
there is a convenience store next door.Two food courts that have huge variety and
restaurants in the Casino to suit any taste & wallet.... are less than 10 min walk
away.
```

```
The staff from Front Desk to cleaners could not be faulted... friendly and help-
ful, making us feel like welcome guests.
```

```
Booked by work for it's location this was a rather expensive XXX YYY stay. Wifi
was provided by an external provider with very expensive rates. This is not great
for people on business. The room was nothing special with a standard shower and
mediocre bed. Clean but pretty bog standard. Nothing to rave about and equally
nothing terrible to report.
```

Fig. 1. Example of reviews containing poorly structured content, subjective sentiments and opinions, metaphor, or cultural references.

In this context, sentiment analysis must play a major role when summarising reviews. Sentiment analysis task can be decomposed in several steps. As a first step, analysis of small texts (phrases, tweets, SMS messages) gives the trend of the conveyed sentiment (commonly refereed to as polarity) generally classified as: positive, negative or neutral. Further steps are needed to summarise the global sentiments. The main difficulty is to give a fair and non-biased picture of the global feeling emerging from individual sentiments. This global picture can consist in a set of numbers (tables, charts, graph. . .) or in a short text that gives the global sentiment in a concise way.

In this study, we propose to compare three approaches to summarisation in order to draw out their current limitations and advantages for this task. This comparison is described in Section 3. Based on this comparison, we propose in Section 4 a new architecture for review summarisation which relies on an abstractive approach making use of domain knowledge and temporal analysis. We conclude the paper with an description of research directions for improving the efficiency of review summarisation methods.

2 Related Research

Summarising opinions reviews into texts can be done in several ways. The most straightforward being the use of a general summariser. Other approaches proposed to produce a tailored "voice summary" of a set of the most extreme restaurants reviews [2]. In the tailored-summariser ReSum [3] the target are reviews on products sold on-line. ReSum outputs two summaries, one for the positive reviews, one for the negative reviews. These summaries are composed of sentences extracted from the positive (or negative) reviews according to a strategy involving redundancy elimination and domain-feature depend criteria such as technical level or Time of Ownership. Here and in the following,

features refer to domain characteristics. For instance, in the accommodation domain, quality of beds or cleanness of the room are domain features (or aspects). While these approaches provide interesting summaries they do not consider opinions in a systematic way, hence the need for a sentiment analysis module in the summarisation framework.

Sentiment summarisation involves several steps (for a detailed review the reader is referred to [4]). The first step aims at determining the sentiment express by each individual reviews. Representative examples of this step are [5] and [6]. [5] proposes a method for determining opinion polarity using WordNet, SentiWordNet and the General Inquirer (to detect polarity shifter). In [6], The probability $P(+/rv)$ (resp $P(-/rv)$) of a movie review rv of being a review of positive (resp. negative) polarity is estimated through the use of Naive Bayes and Markov Model techniques. Each individual review is then scored and this score is used to retrieve the most extreme reviews. However, the method does not capture the global sentiment emerging from the reviews.

The global sentiment of a set of reviews can be abridged as numbers or charts. For example, [7] summarises hotel reviews through automatic features extractions and polarity measure. For each review, if a feature is identified, its polarity is computed. The global sentiment for each feature is then computed. In [8], reviews are summarised in a similar manner but using a domain ontology for features identification. An important advantages of this approach is that it proposes to highlight positive (reps. negative) comments within negative (resp. positive) reviews arguing that opinions about features are more interesting when extracted from a review containing contrasted opinions.

The next section gives a more detailed focus on "pro" and "cons" associated to three methods for the summarisation of the global sentiment emerging from hotel reviews.

3 Comparative studies of three approaches

Three different approaches used to summarise the overall opinion emerging from a set of hotel reviews are presented. The reviews were all collected from the Tripadvisor website. The first experiment concerns the use of a general summariser. The second one shows results obtained when sentences extraction is guided by domain features. The third one consists in the Reviews and Ratings (RnR) system described in [8].

3.1 Open Text Summarizer

Open Text Summarizer (OTS) is an open source tool for summarising texts of any domain [9]. Its content selection is based on the TF-IDF measure with some re-weighting based on the structure of the document (e.g., title and paragraph). The experiment with OTS consisted in feeding it with a whole set of reviews about the same hotel and checking the output. Figure 2 shows an output when OTS was applied with a 1% compression ratio. It can be noticed that no relevant information about the hotel appears before the fourth sentences. As with any extractive summariser, some referring expressions are impossible to understand (e.g.,"**This** appeared from the unlocked. . . "). Moreover, there is no way for the summariser to filter out irrelevant information for the decision making task such as with information about booking experience (e.g., "booking was done at very last minute. . . ", "I did a lot of research. . . "). This is due to the high frequency of personal booking experiences that biased the system towards this kind of irrelevant information. It appears from this short example, that purely frequency-based content

selection without the involvement of some domain and/or task knowledge is unpromising.

```
Hotel booking was done at very last minute by the friendly staff at the
International Airport. I did a lot of research in advance - most of it on
Tripadvisor - and it was ranked very highly. This appeared from the unlocked
office behind reception - I was told this was more secure - I wondered but all was
ok. Location is what this hotel has going for it - you're on holidays, you want to
be in the centre of things, near good restaurants [...]
```

Fig. 2. Beginning of the output of OTS at 1% compression rate (2500 to 17 sentences).

3.2 Features-based Selection Extraction

In this approach, the main idea is to extract relevant information related to a particular word. In [10], an approach to better understand the particular meaning associated to a word in the mind of a particular author was proposed. We proposed to use this technique to capture the global opinion given by a set of users on a particular domain feature.

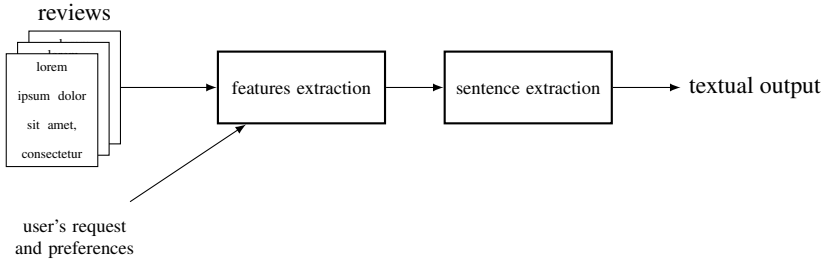

Fig. 3. Diagram of the extractive system

Figure 3 shows the outline of the proposed method. Let denote by f a word type representing a particular feature under study (BED for example). The set of sentences containing f is a sub-corpus denoted by U_f and is called the *lexical universe of* f. For each word type i of the global corpus (C) under consideration two frequencies can be observed:

- $p_{U_f}(i)$ is the observed frequency of type i in U_f the lexical universe of f.
- $p_C(i)$ is the frequency of type i in the whole corpus.

Using the hypergeometric law, an expected value $E_{U_f}(i)$ for $p_{U_f}(i)$ can be computed. Given a confidence level (5% or 1%) it is then possible to tell if the observed value $p_{U_f}(i)$ is too *far* from $E_{U_f}(i)$, either because $p_{U_f}(i) << E_{U_f}(i)$ or because $p_{U_f}(i) >> E_{U_f}(i)$. So each word type i of the whole corpus C can be classified, with regards to f as being:

– *neutral*, this is a set of words that do not have a special interaction with f. There frequency in the lexical universe of f is acceptable with regards to their frequency in the whole corpus;
– *attracted* (if $p_{U_f}(i) >> E_{U_f}(i)$), this is the set of words that are over-represented in the lexical universe of f. They can be seen as being attracted by f and it can be inferred that they are characterizing the global opinion on f;
– *repulsed* (if $p_{U_f}(i) << E_{U_f}(i)$), this is the set of words that are under-represented in the lexical universe of f. They can be seen as being repulsed by f and it can be inferred that they are not reflecting the global opinion on f;

Given a feature f it is then possible to build two sets. U_f^+ the set of words that mostly *characterize* f and U_f^- the set of words that are mostly *repulsed* by f. These sets are used to score each sentences of the whole corpus C so to select the set of sentences that characterize the best the opinion associated to a particular feature.

Figure 4 shows the most relevant sentences for the feature BED. It can be noticed that the most relevant and condense sentences are the best rated. However, there is a high redundancy in this list and contrasted reviews are not fetched by the method.

0.647 A comfortable double bed, couch and coffee table, plus a small desk with two chairs.
0.615 The room was spacious with a queen sized bed and a sofa bed.
0.412 The hotel rooms were a good size with a double bed and a fould out sofa bed.
0.378 Queen sized bed (with small side shelfs), little couch and coffee table for persons, a basic table with chairs, a flat screen tv, and a dresser with a couple of drawers.
0.370 The rooms were quite large - we had a queen room which consisted of a queen bed, small lounge, small table and chairs and kitchenette.
0.364 The room was quite large with a couch, desk and amp ; coffee table as well as the queen size bed.
...

Fig. 4. Example of extracted sentences for the feature BED.

3.3 RnR system

In [8], an RnR system [1] for extracting rationale from on-line reviews/ratings is presented. The system captures and summarises the key rationale for positive and negative opinions expressed in a corpus of reviews and highlights the negative features among positive reviews and vice versa. One of the main contribution of the work is the techniques that have been designed to leverage support metric in conjunction with a domain ontology. This results in improved computational overheads associated with sentiment identification. In term of presentation, the system outputs the summary for each hotel in a four-quarter screen presented in Figure 5. The top left quarter shows the general/summarised overview of the hotel, top right column contains the time based performance chart, and the two bottom sections give details of each positive (left hand side) and negative (right hand side) groups of reviews.

[1] The RnR system is accessible at http://rnrsystem.com/RnRSystem

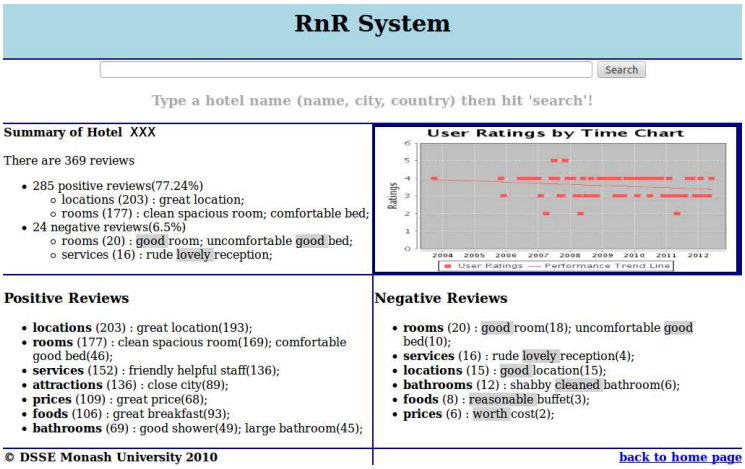

Fig. 5. RnR output.

Though the RnR output provides the useful global picture of the reviews, it is lacking a fundamental dimension which is the temporal dimension. The rating chart does indeed give trends but is of little interest when the tend is flat as it the case in Figure 5. So there is no way for the costumer to know the latest positive and negative features of the hostel nor to know what are the positive and negative constants of it. Furthermore, tabular and keyword presentation might not be the best way of presenting a summarisation of the reviews as every pieces of information is presented in an out-of-context way. A more elegant approach to present such information both with respect to the temporal and contextual perspectives is to use Natural Language Generation (NLG).

4 Towards an abstractive summarisation system

NLG systems has been used for decade to present numerical and linguistic information in a condense and efficient way. Recently, NLG has been applied to summarise large volumes of heterogeneous temporal data to short texts in the medical domain [11]. This system was experimented at the hospital and has shown that a textual-only output can led to better decision from the medical staff than a classical graphical-only presentation. Among the properties, emphasised by the authors [12], that textual summarisation offers compared with the graphical presentation are : the capacity to present data in the same sentence at multiple time resolution or period (e.g., “the hotel had always been praised for its good beds”, “in summer, the hotel is found to be badly ventilated”), the natural ability to handle vagueness and uncertainty (e.g., “the hotel seems to be close to public transport”), the capacity to insert genuine citations (e.g., “the hotel could not even offer us a hand towel!”), the possibility to aggregate features (e.g., “close station(90); free tram(44); close train(33);” $\rightarrow$ “close public transport and free tram”) and the capacity to contrast features (e.g., “even the negatives reviews reports that the bathroom is generally clean and large”).

To address the above limitations and progress beyond the state of the art in this domain, we plan to build an approach based on the work of Rahayu *et al.* [8] and Portet *at al.* [11]. This approach combines a sentiment analyses and domain-specific text processing approaches to represent the data in a high level representation (e.g., in the form of an ontology) with a natural language generation system to generate a textual user-tailored review of an hostel. The intended system is depicted Figure 6. User requests a summary of a specific hostel. In some cases, she can also specify which features are the most important for her so that features belonging to her preferences are given more weight. The system then fetches all the opinions about this hotel (e.g., trip advisor) and extract the features describing each reviews. Once the features extractions is performed, a sentiment analysis layer extracts polarity affecting each phrases of interest. These phrases are then abstracted into facts in a database backed by an ontology which represent the hotel and customer's concepts. Using the ratings, a time series segmentation [13] is performed to identify the main periods of the hotel (decrease, increase, stable). Another segmentation is performed at the feature level to detect specific evolutions of the hotel's services. Once the opinions have been analysed all the data is summarised through an NLG approach.

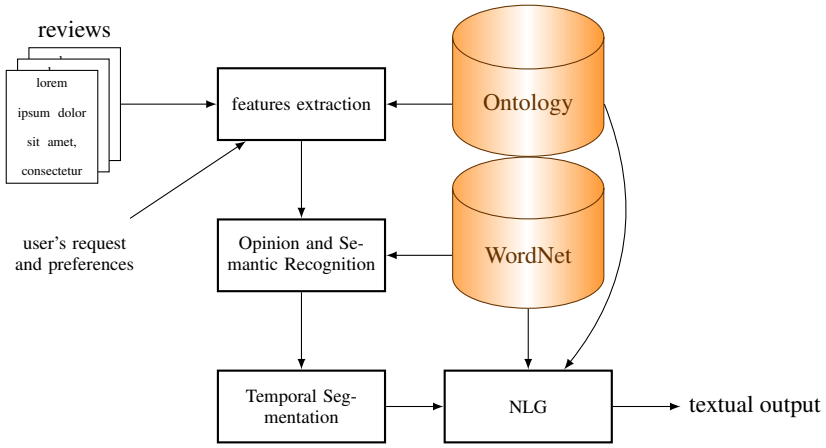

Fig. 6. Diagram of the abstractive system

5 Conclusion

Although human summaries are typically abstracts, most existing systems produce extracts, due to several studies reporting better results of the latter [1]. This is due to the complexity the process that involves concepts extraction, reasoning at the semantic level and natural language generation. This makes it a time consuming task. However, review summarisation is a very different application than documents considered in classical summarisation. The high number of authors, style, subjectivity and the temporal

dimension calls for the reconsideration of the abstractive approaches to perform a deep analysis to better condense the information present in the reviews. Our approach by considering these aspect while aiming for a modular architecture, is a step towards addressing this challenge.

Another important challenge is the evaluation of such technology. This is delicate given that no gold standard summary exists in this domain for automatic scoring (such as with BLEU or ROUGE) and because users will often disagree on what constitutes the best content and quality for the summary. A more relevant measure would be to perform some task-based experiments to assess the effectiveness of the summariser in searching for an hotel. We plan to investigate the techniques used in different domains to propose a formal evaluation strategy which would make it possible to assess the progress of the method.

References

1. Mani, I., Maybury, M., eds.: Advances in Automatic Text Summarization. MIT Press (1999)
2. Mahajan, M., Nguyen, P., Zweig, G.: Summarization of multiple user reviews in the restaurant domain. Technical Report MSR-TR-2007-126, Microsoft Research (2007)
3. Kokkoras, F., Lampridou, E., Ntonas, K., Vlahavas, I.: Summarization of multiple, metadata rich, product reviews. In: Workshop on Mining Social Data (MSoDa), 18th European Conference on Artificial Intelligence (ECAI '08), Patras, Greece (2008)
4. Tang, H., Tan, S., Cheng, X.: A survey on sentiment detection of reviews. Expert Syst. Appl. **36**(7) (September 2009) 10760–10773
5. Martín-Wanton, T., Pons-Porrata, A., Montoyo-Guijarro, A., Balahur, A.: Opinion polarity detection - using word sense disambiguation to determine the polarity of opinions. In Filipe, J., Fred, A.L.N., Sharp, B., eds.: ICAART (1). (2010) 483–486
6. Salvetti, F., Lewis, S., Reichenbach, C.: Automatic opinion polarity classification of movie reviews. Colorado Research in Linguistics **17**(1) (2004)
7. Popescu, A.M., Etzioni, O.: Extracting product features and opinions from reviews. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. HLT '05, Stroudsburg, PA, USA, Association for Computational Linguistics (2005) 339–346
8. Rahayu, D., Krishnaswamy, S., Labbé, C., Alahakoon, O.: Web services for analysing and summarising online opinions and reviews. In: ServiceWave. (2010)
9. Rotem, N.: Open text summarizer (ots) (2003) Retrieved June, 2012, http://libots.sourceforge.net.
10. Labbé, C., Labbé, D.: How to measure the meanings of words? Amour in Corneille's work. Language Resources and Evaluation **39**(4) (2005) 335–351
11. Portet, F., Reiter, E., Gatt, A., Hunter, J., Sripada, S., Freer, Y., Sykes, C.: Automatic generation of textual summaries from neonatal intensive care data. Artificial Intelligence **173**(7–8) (2009) 789–816
12. Gatt, A., Portet, F., Reiter, E., Hunter, J., Mahamood, S., Moncur, W., Sripada, S.: From data to text in the neonatal intensive care unit: Using NLG technology for decision support and information management. AI Communications **22**(3) (2009) 153–186
13. Charbonnier, S., Portet, F.: A self-tuning adaptive trend extraction method for process monitoring and diagnosis. Journal of Process Control **22** (2012) 1127–1138